Análisis de dos sistemas de reconocimiento del habla para su combinación con un editor de subtítulos para programas en directo o sin guion

Trabajo de Fin de Máster

Autora: Andrea Esquer Salom
Tutor: Adrià Martín Mor

Máster en Tradumática: Tecnologías de la Traducción

Facultad de Traducción e Interpretación
Universitat Autònoma de Barcelona

Curso 2018-2019

Resumen

Es cada vez más común el uso de sistemas de reconocimiento del habla para llevar a cabo distintas tareas, como buscar con nuestros dispositivos o dictar texto. En este trabajo se pretende analizar dos sistemas de reconocimiento del habla para su combinación con un editor de subtítulos para programas en directo o sin guion. El principal objetivo es analizar el dictado de dos vídeos, uno en español y otro en inglés, y los errores que aparecen tras este dictado. Para ello se han usado los sistemas de reconocimiento del habla *Dragon NaturallySpeaking* y *Windows Speech Recognition*, con un entrenamiento de estos básico, y se han dictado dos textos extraídos de programas de entrevistas de un minuto de duración. Tras realizar el análisis de los resultados obtenidos podemos afirmar que la mayoría de errores que cometen los sistemas de reconocimiento del habla son de reconocimiento de algunos términos que pueden afectar a la comprensión del texto. En conclusión, para que los sistemas funcionen de manera correcta se deben tener en cuenta diversos factores, como un buen entrenamiento. Y, en el caso de su combinación con un editor de subtítulos también se debe prestar atención a todo aquello relacionado con la subtitulación, como por ejemplo el pautaje del texto.

Palabras clave: Sistema de reconocimiento del habla, *Dragon NaturallySpeaking*, *Windows Speech Recognition*, subtitulación, subtítulos en directo.

Abstract

It is increasingly common to use speech recognition systems to carry out different tasks, such as searching with our devices or dictating text. In this paper, we intend to analyse two speech recognition systems for their combination with a subtitle editor for live or non-scripted programs. The main objective is to analyse the dictation of two videos, one in Spanish and one in English, and the errors that appear after this dictation. To this end, speech recognition systems Dragon NaturallySpeaking and Windows Speech Recognition have been used, with a basic training of these, and two texts extracted from one-minute talk programs have been dictated. After carrying out the analysis of the obtained results, we can affirm that the majority of errors that the speech recognition systems make are recognition of some terms that may affect the comprehension of the text. In conclusion, for systems to function correctly, several factors must be taken into account, such as good training. And, in the case of its

combination with a subtitle editor, attention should be also paid to everything related to subtitling, such as for example the parameters of the text.

Key words: Speech recognition, *Dragon NaturallySpeaking*, *Windows Speech Recognition*, subtitling, live subtitles.

Índice

1. Objetivos

El objetivo de este trabajo consiste en analizar la eficacia de dos sistemas de reconocimiento del habla para su combinación con un editor de subtítulos para programas en directo o sin guion, en español e inglés. Lo que se pretende es ver cuáles son los errores más frecuentes que se producen al usar un sistema de reconocimiento del habla, usando los sistemas *Dragon* y *Windows Speech Recognition* mediante los que se dictará un texto de dos programas de los que no se tiene el guion para comprobar su eficacia, en dos idiomas distintos; y, tras esto, ver que mejoras se pueden aplicar para estos sistemas según los errores que se hayan encontrado.

2. Marco teórico

2.1 Sistemas de reconocimiento del habla

Uno de los componentes que se va a analizar en este trabajo son los sistemas de reconocimiento del habla. Un sistema de reconocimiento de habla (ASR, *Automatic Speech Recognition*) consiste en un software que reconoce las palabras habladas; lo que hacen es comparar las muestras de las palabras habladas con las almacenadas en una memoria, es decir, con las almacenadas en un vocabulario. (Raud et Brennan, 2000) A continuación, se explica cómo surgieron, como funcionan y cuáles son los más conocidos.

2.1.1. Historia de los sistemas de reconocimiento del habla

A lo largo de los años, los sistemas de reconocimiento del habla se han ido mejorando hasta llegar a los sistemas que tenemos hoy en día. Fue en la década de los 50 cuando se empezaron a *desarrollar* los primeros sistemas de reconocimiento del habla, que reconocían un vocabulario reducido, del orden de 10 palabras y emitidas por un único locutor. Ya en los 60 fue cuando se desarrollaron los proyectos con mayor repercusión en esta área: normalización de la voz, para que reconociera más de un locutor; aplicación de métodos de

programación dinámica y primeros trabajos en reconocimiento de voz continua, en lugar de la construcción de 10 palabras cada vez. En los 70 se empiezan a reconocer palabras aisladas y los sistemas se basan en programación dinámica; pero, en los 80 los sistemas se basan en los modelos ocultos de Markov, un modelo estadístico, y ya se empezaron a usar algunas aproximaciones a las redes neuronales. (Bueno et. al., 2007)

A día de hoy, la mayoría de sistemas se basan o bien en los modelos ocultos de Markov o bien de manera híbrida con redes neuronales y pueden llegar a un reconocimiento del orden del 95%, siempre que se trate de un único hablante, con un micrófono de buena calidad y en un ambiente de bajo ruido. (Bueno et. al., 2007)

2.1.2. Componentes y fases en un sistema de reconocimiento del habla y aplicación

Un sistema de reconocimiento del habla está compuesto principalmente por los siguientes modelos básicos de habla: modelo acústico, gramática y vocabulario y modelo lingüístico. Funciona al igual que un sistema de traducción automática estadística, combinando los distintos modelos (Romero-Fresco, 2011).

El modelo acústico es la base de datos que incluye el audio, su transcripción, corregida y revisada, y la representación digital de este audio, es decir, como se representa matemáticamente. La gramática y el vocabulario son un listado de palabras que incluye el sistema para poder reconocer la equivalencia más correcta a la expresión que ha usado el hablante; si esa expresión no está en la lista, puede que reproduzca una opción que no toca o la omita. Y en cuanto al modelo de lengua, se trata de un mecanismo probabilístico que calcula la probabilidad de uso de una cadena de palabras; es decir, calcula la probabilidad del uso de una palabra analizando la usada antes y la siguiente que se usa (Romero-Fresco, 2011).

Las fases del proceso que usa un sistema de reconocimiento del habla son las siguientes: paso de sonido análogo a fonemas, de fonemas a palabras y de palabras a frases. Para pasar el sonido a fonemas, o *front-end phase* (Bueno et. al., 2013) el sistema reduce el ruido de fondo y el volumen del sonido que adquiere y analiza la voz para pasarla a fonemas. A continuación,

se comprueban, mediante probabilidades, los fonemas con la lista de palabras disponibles para elegir las que sean más parecidas y más precisas respecto al sonido que ha reconocido el sistema. Y, finalmente, se calcula la probabilidad de que una palabra dictada se use tras o ante otra; por ejemplo, se calcula la probabilidad de que se haya usado *city* o *sid* tras *new* y *york*. Este último paso se calcula en n-gramas. Si el tema de la base de datos y el que está tratando el hablante es el mismo mejorará la precisión del texto que cree el sistema.

Los sistemas de reconocimiento del habla se pueden aplicar a diferentes ámbitos como el control de teléfonos u otros dispositivos electrónicos o la creación de subtítulos para personas con alguna dificultad auditiva, que es el caso en el que se centrará este trabajo.

2.1.3. Ejemplos de sistemas de reconocimiento del habla

Entre los sistemas más destacados de reconocimiento del habla se encuentra *Dragon*. Su primera versión salió al mercado en 1997, actualmente va por la versión número 15 y está disponible en 8 idiomas y para dos sistemas operativos: Windows y MacOS[1]. *Dragon* ofrece distintas opciones de las diferentes versiones que tiene como por ejemplo la versión *stadard* y la *professional*. Pero también da opciones específicas según algunos temas concretos como *Dragon Medical*[2] o *Dragon Law Enforcement*[3]. Estas últimas opciones ayudarán a mejorar la precisión del sistema cuando se quieran tratar alguno de estos temas, dado que la base de datos se centrará de manera más precisa en estos temas.

Otro sistema de reconocimiento del habla conocido es el *Windows Speech Recognition*. Microsoft lleva desde 1993 trabajando en este sistema, pero su primera versión se incorporó en la versión de Windows Vista, en 2006. Mediante el uso de este sistema podemos controlar nuestro equipo y dictar en alguna de las aplicaciones.

Además de estos dos sistemas, que son propietarios, también podemos encontrar otros gratuitos como *Web Captioner*, que funciona de manera online, entre muchos otros. Por otro lado, encontramos el proyecto *Common Voice* de

[1] https://www.nuance.com/es-es/dragon.html
[2] https://www.nuance.com/es-es/healthcare/physician-and-clinical-speech/dragon-medical.html
[3] https://www.nuance.com/dragon/industry/dragon-law-enforcement.html

Mozilla Firefox, que consiste en la creación de un corpus oral y que de momento sigue en fase de desarrollo aumentando su base de datos sobre distintos temas y en distintos idiomas.

2.2. La subtitulación

La subtitulación, como explica Chaume (2003), es la introducción de texto en pantalla donde se proyecta un contenido visual, haciendo que coincida de manera aproximada con las intervenciones de los interlocutores que aparecen. Según Bartoll (2015), la subtitulación presenta una serie de características que la determinan, y estas pueden ser textuales o formales.

Las características textuales son el cambio de canal, de oral a escrito, y la relación entre texto e imagen. Al hacer un cambio de canal, se deberá tener en cuenta que se perderá información, como la entonación o el acento de algún interlocutor, en caso de que sea un aspecto relevante. Dado que la lengua escrita es más limitada que la oral, deberemos reflejar toda aquella información que es relevante en los subtítulos. En cuanto a la relación entre el texto y la imagen, de los aspectos más relevantes que deberemos tener en que, siempre que lo permita el tipo de subtitulación que se está llevando a cabo, que el subtítulo empiece y acabe con la intervención oral del interlocutor que subtitulamos.

En referencia a las características formales, estas marcan los límites de caracteres por línea y el número de líneas por subtítulo. Estas características también tienen en cuenta el tipo de letra, los estilos que se usan y la duración del subtítulo en pantalla. Según el tipo de subtitulación el límite de caracteres puede variar, dado que no es lo mismo subtitular para alguien que puede oír el audio original, que para alguien que tiene alguna discapacidad auditiva.

En traducción audiovisual podemos encontrar tres tipos de subtítulos, según el momento en que se realicen (pantallas diversas):

- Subtítulos en diferido: se realizan antes de emisión del programa.
- Subtítulos en semidirecto: se termina antes de la emisión del programa, pero se sincronizan durante esta.
- Subtítulos en directo: se realizan durante la emisión del programa.

Para subtitular cada plataforma, empresa o soporte tiene su propio proceso de elaboración de subtítulos, pero todos mantienen unas características comunes: la transcripción del original, el pautaje (que cosiste en indicar la duración del subtítulo en pantalla y el tiempo de entrada y salida de este), la traducción si se requiere, la revisión de los subtítulos y su aplicación al material audiovisual.

Respecto a los distintos editores de subtítulos, cada traductor o empresa decide que editor usar para realizar su trabajo. Entre los más conocidos encontramos *Subtitle Workshop*, *Media Subtitler* y *Aegisub*. Los tres editores son de software libre, pero *Media Subtitler* solo funciona para Windows. Todos tienen una configuración y una interfaz parecidas: un visor de video, un editor de texto donde insertar los subtítulos, una lista con los subtítulos creados y sus tiempos de entrada y salida, una zona donde modificar y visualizar el tiempo de los subtítulos, etc.

Gracias a la subtitulación podemos acercar los productos audiovisuales a personas con alguna deficiencia auditiva o, como suele pasar, a personas que no entienden la lengua original. Aunque hay muchos tipos de traducción, en este trabajo analizaremos cómo funciona la subtitulación en directo y, más concretamente, cuando la transcripción del original se lleva a cabo mediante un sistema de reconocimiento de voz y el rehablado.

En los siguientes apartados se explicará cómo funciona la subtitulación en directo, el rehablado y la combinación de un editor de subtítulos con un sistema de reconocimiento del habla.

2.3. Subtitulación de programas en directo o sin guion y rehablado

Como ya hemos explicado en el apartado anterior, la subtitulación juega un papel muy importante en lo que se conoce como accesibilidad del contenido multimedia. Es por esto por lo que, para poder hablar de subtitulación en directo, debemos tener en cuenta que esta práctica se suele llevar a cabo para que las personas sordas tengan acceso al contenido multimedia. La subtitulación en directo, que también se conoce como subtitulación simultánea (Rodríguez, 2005), consiste en ir insertando los subtítulos a la vez que se emite el programa.

Como explica Romero-Fresco (2011) el rehablado se podría definir como la producción de subtítulos mediante el uso de un sistema de reconocimiento de voz, aunque como definición conceptual añade la siguiente en su libro:

> A technique in which a respeaker listens to the original sound of a live programme or event and respeaks it, including punctuation marks and some specific features for the deaf and hard of hearing audience, to a speech recognition software, which turns the recognized utterances into subtitles displayed on the screen with the shortest possible delay.

A pesar de que el rehablado se usa de manera habitual para la creación de subtítulos en directo, también puede usarse para la creación de subtítulos pregrabados, dado que facilita la transcripción de estos cuando el traductor no tiene acceso al guion del producto audiovisual que debe traducir. Pero, como ya se ha mencionado su uso más común es para la subtitulación en directo, de la que se explica a continuación como empezó a usarse en TV.

En 1990, la BBC fue la primera cadena en crear una unidad de subtitulado en directo y comenzó a usar el rehablado en 2001 (Marsh, 2006; en Lago 2013). Aquí en España, la primera cadena en introducir la subtitulación en directo fue TV3, mediante un sistema de semáforos para subtitular. RTVE también empezó más o menos a la vez que la cadena catalana. El sistema, preparado por TV3 era para cinco personas con teclados QWERTY, funcionaba de la siguiente manera: cuando en la pantalla aparecía la franja de color en verde, la persona encargada de introducir los subtítulos escribía lo que oía por los auriculares y cuando aparecía la franja roja, dejaba de escribir. Aunque la mayoría de programas tenían subtítulos en semi-directo, es decir, que una parte de estos ya estaba introducida antes de que el programa se emitiera, había dos programas que sí que tenían los subtítulos en directo: los partidos de fútbol y el programa *Àgora* en 2005. En total, TV3 producía 160 horas de subtítulos en directo a la semana (Orero, 2006).

Ilustración 1:Real-time subtitling multiplexing five operators with QWERTY. (©Rosa Vallverdú en Orero, 2006)

Como bien explica González Lago en su artículo, la subtitulación en directo usando un sistema de reconocimiento del habla consta de tres fases:

- El rehablador dicta el texto que escucha a un sistema de reconocimiento del habla.
- Se corrigen los posibles errores del texto que el sistema ha reconocido de forma incorrecta.
- Se emiten los subtítulos mediante otro programa, específico para la emisión de estos.

Este tipo de experiencias suelen llevarse a cabo en pareja. Mientras una de las dos personas dicta el contenido, la otra se centra en corregir y editar los errores que se han transcrito y lanzar los subtítulos al directo. También hay ocasiones en que trabajar en pareja no es posible y la persona que dicta debe corregir y lanzar a su vez los subtítulos.

Además de las distintas fases, se debe tener en cuenta que para la subtitulación para personas sordas existe la Norma UNE 153010[4], cuya última versión es del 2012, que establece los criterios mínimos de calidad y de homogeneidad de este tipo de subtitulación.

En el artículo de González Lago (2013) explica cómo se aplica el modelo NERD para calcular la precisión de los subtítulos en directo, un modelo que explica Romero-Fresco en su libro (2011:150-161). El modelo se basa en

[4] https://www.une.org/encuentra-tu-norma/busca-tu-norma/norma/?c=N0049426

una fórmula que al número de palabras del texto rehablado más los signos de puntuación y comandos usados (N), se le restan los errores de edición (E) , los errores de reconocimiento (R) y los errores leves (tildes, mayúsculas, etc.)(D) se divide entre el número de palabras del texto rehablado (N) y se multiplica por 100. El resultado será el porcentaje de precisión de esos subtítulos.

2.4. Combinación de un sistema de reconocimiento del habla y un editor de subtítulos

A continuación, se explicarán dos proyectos en los que se combinan ambos sistemas.

2.4.1. Proyecto VOICE

El proyecto VOICE (2004) consiste en una investigación del uso de los sistemas de reconocimiento de voz para conversaciones, conferencias, televisión y llamadas telefónicas. Su principal objetivo es acercar a las personas con alguna discapacidad auditiva todos los productos auditivos que puedan mediante un sistema de reconocimiento de la voz combinado con otro sistema, según de que producto se trate. Comenzó en 1996 por *EC Joint Research Centre* en *Inspra* y de 1998 a 2002 pasó a financiarlo *EC Directorate General Information Society.*

Fue entre los años 2001 y 2002 en los que el proyecto se centró en la subtitulación. Según la web del proyecto, la aplicación de un sistema de reconocimiento del habla facilitaría muchos aspectos de la subtitulación, como el coste o el aumento de material audiovisual que tendría subtítulos.

Esta es la imagen que muestran como ejemplo de cómo sería el software para subtitulado:

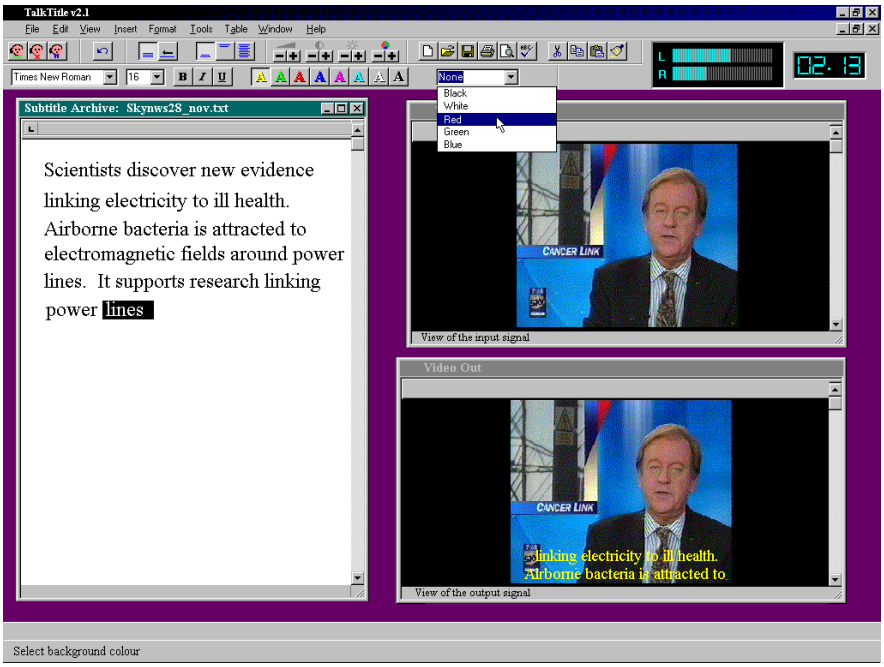

Ilustración 2: Proyecto VOICE, interfaz de subtítulos (2004)

2.4.2. Proyecto MUSA

El proyecto MUSA (Multilingual Subtitling of Multimedia Content), que se llevó a cabo entre 2002 y 2005, fue un proyecto desarrollado por el Institute for Language and Speech Processing en Grecia que apuntaba a la creación de un sistema multimodal multilingüe que convierte corrientes de audio en transcripciones de texto, genera subtítulos de estas transcripciones y luego los traduce a otros idiomas. (MUSA, 2004) Los idiomas en los que está disponible este proyecto son el inglés, el francés y el griego.

Lo que pretende el proyecto es combinar un sistema de reconocimiento del habla con una máquina de traducción, como ellos lo denominan, que combine un motor de traducción automática con una memoria de traducción y un módulo de sustitución de términos. Para generar los subtítulos, como explican en su web, se realiza mediante el análisis automático de la estructura lingüística de la oración.

Su objetivo principal es desarrollar un sistema que combine el análisis de texto avanzado, el reconocimiento de voz, la traducción automática y otras técnicas para ayudar en la preparación de subtítulos; todo esto cumpliendo los requerimientos específicos espacio-temporales del proceso de subtitulado. (Bueno et. al., 2007).

Entre sus otros objetivos también se encuentra la combinación de un sistema de reconocimiento del habla con la condensación de los subtítulos y la traducción en un marco unificado que pueda producir buenos subtítulos tanto a nivel interlingüístico como intralingüístico. Dadas las restricciones espacio-temporales que supone la subtitulación, el sistema de subtitulación tendrá en cuenta la transcripción del original para decidir qué se debe incluir en el texto meta y que se puede omitir.

3. Metodología

Para llevar a cabo el análisis de *Dragon* y de *Windows Speech Recognition*, se va a introducir el mismo texto en los dos sistemas de reconocimiento para ver que errores se cometen de manera más frecuente y ver que mejoras se pueden aplicar para que esto no suceda. Ambos sistemas se van a usar simplemente con el entrenamiento básico de estos; es decir, no se han programado para que reconozcan mejor los textos que vamos a introducir o la voz que va a dictar dichos textos, simplemente se hará el entrenamiento obligatorio del programa cuando este se inicia.

El texto que se va a usar para el análisis en español es una parte de la entrevista que realiza Risto Mejide a Guillermo Bárcenas en su programa Chester (Mediaset España, 2019). A pesar de que se trata de un programa de entrevistas, en que el presentador lleva las preguntas preparadas, y no es en estricto directo, no sabemos qué puede contestar el interlocutor, por lo que no se tiene un guion escrito. El vídeo dura alrededor de 7 minutos, pero solo se ha analizado un minuto.

El vídeo elegido para el análisis en inglés es un sketch del programa *Jimmy Kimmel Live!: Lie Detective* (Jimmy Kimmel Live, 2012). En los sketches el presentador del programa, Jimmy Kimmel somete a niños a un detector de mentiras falso. Al igual que la entrevista que se ha elegido para el texto en español, en este podemos saber que preguntas va a formularles a los niños el presentador, pero no qué van a responder a esas preguntas; por tanto, no tenemos un guion del programa, aunque este no sea en directo. Este video dura alrededor de cuatro minutos, pero al igual que el anterior solo se ha analizado uno.

La elección de estos dos vídeos no ha sido del todo aleatoria. Buscamos dos vídeos en los que el componente principal fuera una conversación espontánea, como por ejemplo una entrevista que es el caso de estos dos vídeos. A pesar de que ambos son programas que no se emiten en directo, los dos son sin guion, menos para las preguntas, que pueden tenerlas por escrito. Es por eso por lo que, todo el parlamento debe ser transcrito, no lo encontraremos por escrito en ningún lugar. Además, para elegir el minuto a analizar se han tenido en cuenta diversos factores, según el idioma. Para el español, se ha tenido en cuenta que en esa entrevista apareciesen nombres propios que no son comunes, como el caso Gürtel; el uso de nombres propios que pueden ser también sustantivos comunes y el uso de acrónimos. En cambio, para el inglés se ha tenido en cuenta el uso de contracciones y coloquialismos.

Para analizar los textos que se van a redactar, se usarán los sistemas de reconocimiento del habla y se introducirán en Word. Desde ahí se clasificarán todos aquellos errores que se encuentren según la clasificación que se explicarán en el siguiente apartado. Lo que no se podrá analizar es el tiempo de retardo del subtítulo, es decir, lo que tarda en aparecer el subtítulo en pantalla tras el discurso del interlocutor al que corresponde ese subtítulo. Ni tampoco si el subtítulo es demasiado largo o no. Simplemente se analizará lingüísticamente la reproducción del texto oral a texto escrito.

A continuación se explica cómo funcionan los dos sistemas de reconocimiento del habla y la categorización de los errores.

3.1. *Dragon NaturallySpeaking*

Al iniciar el sistema *Dragon NaturallySpeaking*, obliga al usuario a crearse un perfil para que sea de uso exclusivo y el sistema pueda reconocer mejor la voz y la entonación de la persona que está usando el sistema de reconocimiento de voz. A continuación, se debe completar un entrenamiento, en el que el software permite al usuario elegir el texto que va a dictar entre una selección. Una vez completado este entrenamiento, enseña un aprendizaje interactivo sobre cómo funciona el sistema y permite al usuario familiarizarse con los comandos y demás instrucciones que se van a usar durante el dictado. También, ofrece un menú lateral, en el que aparecen los comandos más

comunes clasificados según el programa o sección del ordenador sobre el que vamos a dictar. Ej.: dar formato, obtener ayuda, etc. Entre las muchas opciones que da *Dragon*, una de ellas es introducir vocabulario ya sea mediante documentos del equipo, emails mandados o editar nosotros mismos el glosario que incorpora de base para mejorar el sistema para cuando dictemos nuestros textos.

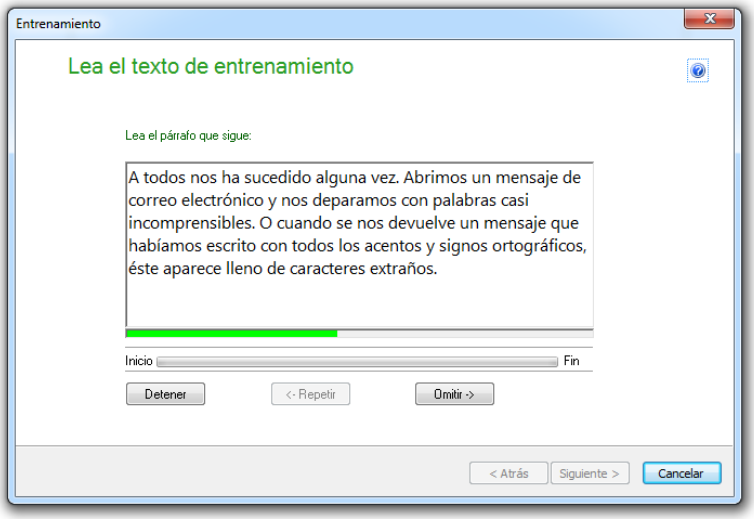

Ilustración 3: Entrenamiento Dragon. Captura de Pantalla, Mayo 2019

3.2. *Windows Speech Recognition*

Cuando empezamos a usar el Sistema de reconocimiento de voz de Windows, al igual que *Dragon*, nos obliga a hacer un entrenamiento previo a usar el sistema; pero este será de una manera o de otra según la versión de Windows con la que estemos trabajando. En el caso de este trabajo los resultados obtenidos son los que se realizaron usando el *Windows Speech Recognition* de Windows 10, en el que el entrenamiento es básico y simplemente se dicta el texto que aparece en pantalla. En cambio, en el caso de Windows 7, el entrenamiento consiste en un aprendizaje interactivo del sistema. A continuación, adjunto dos imágenes de los dos casos:

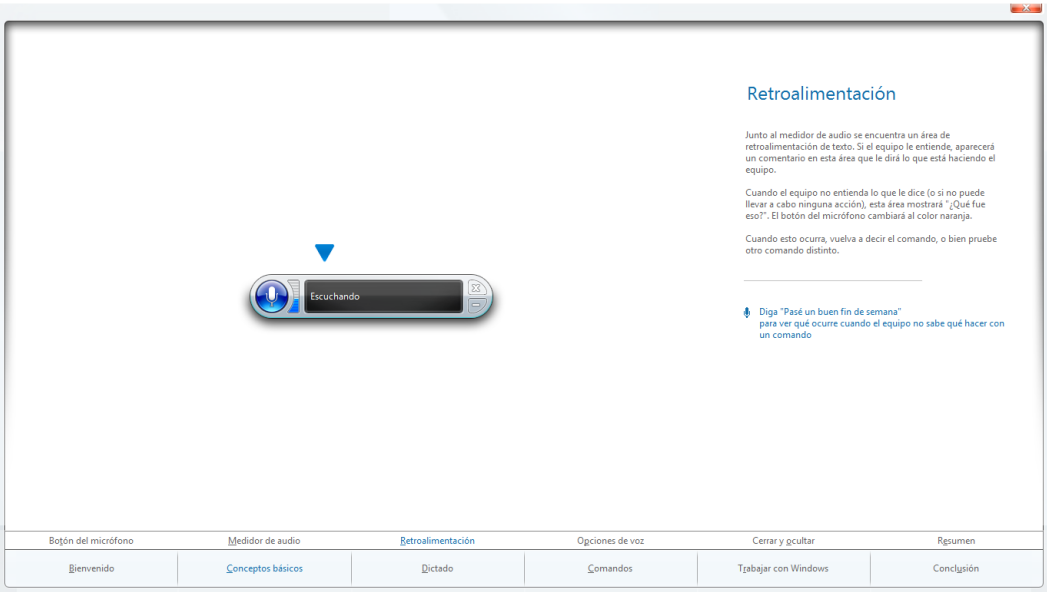

Ilustración 4: Entrenamiento Windows Speech Recognition en Windows 7. Captura de pantalla, Mayo 2019

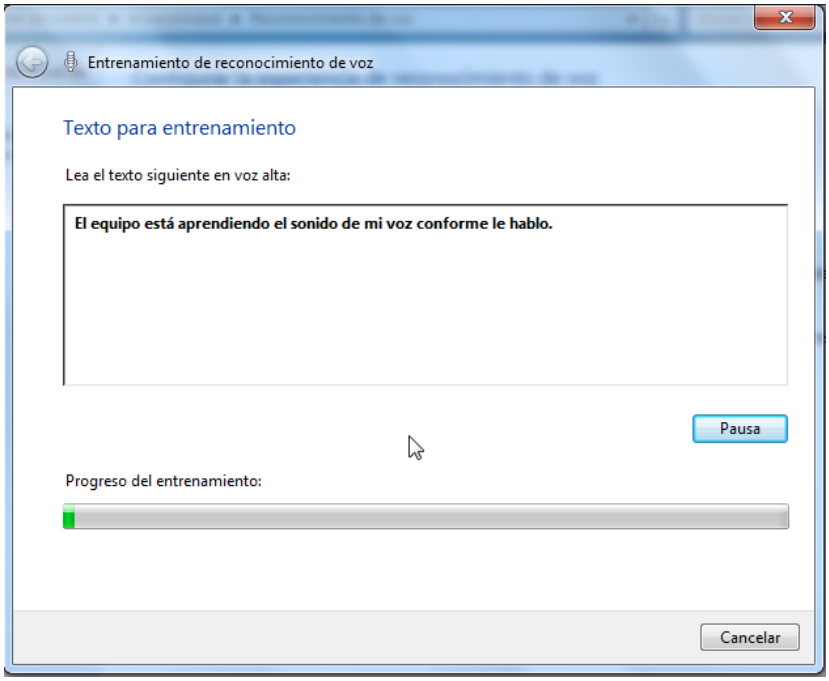

Ilustración 5: Entrenamiento Windows Speech Recognition en Windows Vista y 10. Captura de pantalla, Mayo 2019

3.3. Categorización de errores más frecuentes

La clasificación de errores que se va a usar se basa en parte a la que lleva a cabo Lago (2013) en su análisis a los programas de RTVE, sin tener en cuenta los errores de edición, dado que el análisis se lleva a cabo mediante un texto que nosotros mismo hemos introducido. Por un lado, tendremos los

errores de reconocimiento, que pueden ser adiciones, eliminaciones o sustituciones que ha hecho el sistema de reconocimiento del habla. Por otro lado, errores de ortografía y puntuación, como tildes o mayúsculas o cambios en la puntuación de las oraciones. Y, finalmente, errores graves, que son aquellos que cambian el significado o el sentido del discurso. En el texto, los errores se marcarán con colores, en verde los errores de reconocimiento, en naranja los errores de ortografía y puntuación y en rojo los errores graves.

4. Resultados

A continuación, se analizará cada sistema por separado en cada idioma y se anotarán los errores en cada caso. Y, finalmente, se mencionarán una serie de mejoras para los sistemas de reconocimiento del habla.

4.1. *Dragon NaturallySpeaking*

Tras seguir todos los pasos, y una vez que el programa ha analizado el dictado del usuario, hemos empezado con el dictado del vídeo en español con *Dragon*. Los resultados de este son los siguientes:

Texto original	Texto dictado
ð ¿En qué momento te enteras tú de la que se lía con tu padre? ¿En el momento en el que sale en la prensa o te enteras antes? ð La primera vez que me entero estaba yo en la universidad, en la Francisco de Vitoria, estaba yo acabando ADE. Estábamos estudiando en la biblioteca y recuerdo que vino un amigo y me dijo: ñ S a l e t u p a d r e e n *El Periódico*ò . Y f u i a l l ² , e hace la de Dios. Y esto eran los inicios de la Gürtel. Imagínate lo tonto que soy, que hasta me hizo ilusión.	ð En qué momento te enteras tú de la que seguía con tu padre. En el momento en el que sale en la prensa o te enteras antes. ð La primera vez que me entero estaba yo la Universidad, en la Francisco de vitoria, estaba yo acabando a de. Estamos estudiando la biblioteca y recuerdo que vino un amigo y me dijo: sal de tu padre la portada de el periódico. Y fue allí extraer al año 2009, hace la de Dios. Y esto eran los inicios de la Walter. Imagínate lo tonto que soy, que hasta me hizo ilusión. ð En serio ð no no. A ver, no era ninguna noticia en

ð ¿En serio? ð No, no. A ver, no era ninguna noticia en especial. Pero una ilusión, dije: ¡Coño!. Para que analices la gilipollez de la edad. Dije, pues ahí está, esto se quedará en nada. ð Mira el campeón de mi padre ahí. ð Esto se quedará en nada, se quedará en una noticia más y fin. Pasaba un poco de refilón. No es que fuese el titular absoluto. ð ¿Y lo comentaste con tu padre? ð Si, lo comenté. Le dije: he visto que esto no va a ningún lado, esto son difamaciones, eso son tonterías, esto no va a nada.	especial. Pero una ilusión, dije: coño. Para que analices la gilipollez de la edad. Dije, pues ahí está, esto se quedará en nada. ð Mira el campeón de mi padre ahí. ð Esto se quedará en nada, se quedará en una noticia más y fin. Pasaba un poco de refilón. No es que fuese el titular absoluto. ð Y lo comentaste con tu padre ð sí, lo comente. Le dije: he visto que esto no va a ningún lado, estos son difamaciones, eso son tonterías, esto no va nada.

A continuación, analizaremos los errores de este texto, según la clasificación explicada en el anterior apartado:

Errores de reconocimiento

En este caso, encontramos simplemente dos errores. El primer error es la eliminación de la preposición *en* entre *yo* y *la*; el sistema no debe haber reconocido que ahí la preposición era necesaria. Y el segundo, también se trata de la eliminación de la preposición *en,* pero esta vez entre *padre* y *portada*.

Errores de ortografía y puntuación.

Como ocurre con la mayoría de sistemas de reconocimiento del habla, la puntuación debe dictarse también para que aparezca en el texto dictado. Es por esto por lo que, faltan las exclamaciones, las preguntas, etc.

A parte de la puntuación general, debemos destacar que en ocasiones el sistema decide escribir una palabra en mayúscula o en minúscula de manera indistinta. En este texto encontramos que la palabra *universidad* la ha decidido escribir en mayúscula y la palabra

Vitoria en minúscula, cuando ambas van al revés. El caso de *universidad* puede que se deba a que ha creído que se trataba del nombre propio del centro, que en este caso sí que se escribiría en mayúscula (Ej.: Universidad Autónoma de Barcelona). Pero cuando nos referimos a su significado genérico se escribe en minúscula, aunque esté aceptado el uso de mayúsculas si lo creemos una abreviación del nombre propio. En cuanto a la palabra *Vitoria*, se debe escribir en mayúscula porque forma parte del nombre propio de la institución. El sistema lo debe haber reconocido como la forma desusada del sustantivo *victoria* y por eso lo debe haber escrito en minúscula.

El siguiente error a destacar en este apartado es como ha escrito el sistema *El Periódico*. Al igual que pasa con la puntuación general, cuando queremos que un término aparezca en cursiva, esta se debe dictar para que se aplique. Pero aparte de esto, al tratarse de un nombre propio debería ir en mayúsculas. Como ocurre con muchas otras palabras, al tratarse también de una palabra de uso genérico, la ha interpretado como tal y por eso lo ha escrito en minúscula.

Y, finalmente, el sistema ha iniciado dos intervenciones en minúscula. Ambas empiezan con un adverbio, *no* y *sí*. En la mayoría de casos el sistema reconoce que empieza una nueva oración y más si le informas de que empiece una línea nueva a continuación.

Errores graves

En este apartado se comentarán los errores de uno en uno. En primer lugar, encontramos que *Dragon* ha sustituido *se lía* por *seguía*. El uso de este verbo en este caso modifica el significado de la oración y por eso se considera un error grave.

El siguiente error, es la conversión de las siglas *ADE*, referentes al grado de Administración y Dirección de Empresas, por las preposiciones *a de*. Las siglas son un error frecuente en los sistemas de reconocimiento del habla, ya que en muchas ocasiones pueden confundirse con otras construcciones y, dependiendo de cómo se haya entrenado ese sistema, será capaz de reconocerlo como siglas o como otra cosa.

En el siguiente caso, el cambio de tiempo verbal no acontece y, por tanto, la sustitución de *estábamos* por *estamos* se considera error grave ya que cambia el sentido de la oración. A continuación, *Dragon* ha cambiado *sale* por *sal de*, que no tiene sentido en esta oración; puede que se haya producido un error de reconocimiento, pero se considera grave dado que no tiene ningún sentido añadir la preposición en este caso, que no acompaña ni introduce nada.

El siguiente error que encontramos es el cambio de persona con el verbo *ir*, el original dice *fui,* mientras que, tras haberlo dictado, el sistema ha escrito *fue*. Se trata de un error grave porque cambia por completo el significado de la oración, ya que el sujeto deja de ser el interlocutor. En esta misma oración, encontramos como el sistema ha sustituido *esto era el* por *extraer al*. Lo consideramos error grave porque este cambio no acontece en este caso y modifica la oración hasta el punto de que no tenga sentido alguno.

Y, finalmente, el último error que encontramos en este caso es el cambio de *Gürtel* por *Walter*. El sistema no ha reconocido esa palabra porque no aparece en su lista de vocabulario, pero es necesario que se corrija este error porque da nombre propio a una trama judicial y, por tanto, se considera error grave porque la sustitución conlleva que no se reconozca a que se está haciendo referencia.

Para el texto en inglés, el resultado obtenido tras dictarlo en *Dragon NaturallySpeaking* es el siguiente:

Texto original	Texto dictado
ð How are you? ð Good. ð Are you having a fun day? ð Yeah. ð Yeah. il cô emr oJf if mmy, i meet you. Have you ever met a police officer before? ð Yeah. ð Have you ever taken a lie detector before? You have. O	ð how are you ð good ð Are you having a fun day ð yeah ð yeah. il cô emr oJf if mmy, i meet you.have you ever met a police officer before ð yeah ð have you ever taking a lie detector before.you have.okay good, so

gonna be an old handed on this. So, the truth fairy is going to put these things on your fingers. He ôs Now we ôre put this helmet on you. Have you ever worn a helmet before? Like Ironman, right? Yeah you some questions and all you have to do is tell me the truth, ok? ð Yeah. ð You know what the truth is? ð Yeah. ð Ok, perfect then. What is your name? ð Eric. ð Eric, how old are you? ð Four. ð Do you go to school? ð Yeah. ð Ok. Do you like school? ð Yeah.	going to be an all ended on this.so, the truth fairy is going to put this things on your fingers.she is gone hook you up. now they gone put this helmet on you.have you ever worn helmet before? Like I run men, right? Ye gonna ta ask you some questions and all you have to do is tell me the truth, okay ð year ð you know what the truth is ð year ð okay, perfect then.what is your name ð Eric ð Eric, how old are you ð 4. ð do you go to school ð year ð okay.do you like at school ð yeah

Como sucede con el texto en español, al dictar el texto en inglés también aparecen errores.

Errores de reconocimiento.

El primer error que encontramos en este apartado es la eliminación del artículo *a* justo antes del sustantivo *helmet*. No es un error grave, dado que no modifica ni el sentido ni el significado de la oración. Después encontramos hasta en cuatro ocasiones el uso de *okay*. No se trata de un error como tal, pero estamos analizando esta transcripción para aplicarla a unos subtítulos, por tanto, hay que tener en cuenta que cuantos menos caracteres usemos, mejor, ya que tenemos un espacio reducido. Es por eso por lo que, en este caso se podría haber usado la abreviatura *ok*.

El siguiente error que se puede destacar es el uso de un número. No existe ninguna norma escrita en inglés sobre si escribir la cifra con un número o con letras, pero según *The Chicago Manual of Style* debe escribirse en letras desde el número cero hasta el cien, además de

todos aquellos combinados con cien, mil, millones, etc. Como sucede en este caso, *four* debería escribirse en letras y no en números.

Y, por último, *Dragon* ha añadido la preposición *at* en la última oración. No se marca como error grave, a pesar de que cambie el significado de la oración, porque no afecta a la contestación del interlocutor, pero debería cambiarse antes de mantenerlo en el subtítulo.

Errores de ortografía y puntuación.

En el caso de la puntuación, pasa como con el otro texto; si no se dicta, no se escribe. Por tanto, faltan todas las interrogaciones, exclamaciones, etc. Aparte de la puntuación general, podemos destacar en este apartado el uso de *going to* en lugar del *gonna*, coloquial. No es un error, ni se podría considerar como tal, pero teniendo en cuenta que se trata de un programa americano y que el texto que hemos dictado se convertirá en subtítulos, deberemos intentar reducir todo lo posible los caracteres que se van a usar. Además, al tratarse de una construcción típicamente oral, solo se admitiría su uso cuando debemos ajustar los caracteres y estamos haciendo subtitulación intralingüística, que es cuando intentamos mantenernos lo más fieles posibles al audio.

El siguiente error que encontramos se trata de la sustitución de *he's* por *she's*. Es un error de ortografía, *Dragon* ha reconocido *this* pero debería usar su plural, *these*. Y el último error es la sustitución de *he's* por *she's*. Se trata de un error de reconocimiento, pero lo podemos clasificar como error de ortografía porque simplemente ha añadido una letra cuando no tocaba.

Errores graves

En esta categoría de errores encontramos que el sistema ha sustituido *taken* por *taking*, cuando no acontece en este caso y crea un error gramatical. El siguiente caso es un error de reconocimiento que cambia por completo el sentido y el significado de la oración, que es la sustitución de *old handed* por *all ended*.

A continuación, se ha sustituido *going to* por *gone*, que cambia por el tiempo verbal y el sentido a la frase. En el siguiente caso no solo se cambia el tiempo verbal, sino que también se sustituye el sujeto: pasa de *we're gonna* a *they gone*. El cambio en este caso no tiene ningún sentido, ya que cambia por completo el significado y el sentido de la oración.

El siguiente error se produce debido a que el sistema no tiene un vocabulario tan extenso como para recoger todo tipo de nombres propios; como en este caso pasa con *Ironman* que lo sustituye por *I run men*. Dado que pasa de un nombre propio a una construcción sintáctica distinta, se considera error grave porque además cambia el significado de la oración.

Y, finalmente, el último error que detectamos en este texto es el cambio de la expresión *yeah* por *year*. Puede que se deba a que el sistema no está lo suficientemente entrenado como para diferenciar con claridad la pronunciación de una palabra y de otra. Se considera error grave porque cambia tanto el sentido como el significado de la interlocución y no acontece en este caso el uso de la palabra *year*.

4.2. *Windows Speech Recognition*

Tras haber redactado los dos textos mediante el sistema de *Windows Speech Recognition*, estos son los resultados que se han obtenido de cada idioma. En la siguiente tabla aparece el texto original y el texto resultante del dictado del vídeo en español:

Texto original	Texto dictado
ð ¿En qué momento te enteras tú de la que se lía con tu padre? ¿En el momento en el que sale en la prensa o te enteras antes? ð La primera vez que me entero estaba yo en la universidad, en la Francisco de Vitoria, estaba yo acabando ADE.	ð En qué momento te enteras su de la que seguía con tu Padre y en el momento en el que sale en la prensa o te enteras antes ð La primera vez que me entero estaba yo en la universidad, in la Francisco de Vitoria estaba yo acabando ante. Estamos

Estábamos estudiando en la biblioteca y recuerdo que vino un amigo y me dijo: ñ S a l e t u p a d r e e n *El Periódico*ò . Y f u i a l l ² , hace la de Dios. Y esto eran los inicios de la Gürtel. Imagínate lo tonto que soy, que hasta me hizo ilusión. ð ¿En serio? ð No, no. A ver, no era ninguna noticia en especial. Pero una ilusión, dije: ¡Coño!. Para que analices la gilipollez de la edad. Dije, pues ahí está, esto se quedará en nada. ð Mira el campeón de mi padre ahí. ð Esto se quedará en nada, se quedará en una noticia más y fin. Pasaba un poco de refilón. No es que fuese el titular absoluto. ð ¿Y lo comentaste con tu padre? ð Si, lo comenté. Le dije: he visto que esto no va a ningún lado, esto son difamaciones, eso son tonterías, esto no va a nada.	estudiando en la biblioteca y recuerdo que vino un amigo y me dijo: salí tu Padre en la portada del periódico. Y fui allí esto era el año 2009, hace la que Dios, y esto Irán los inicios de muster. Imagínate no tonto que soy que hasta ministro ilusión. ð En serio ð No no abrir no era ninguna noticia en especial. Tiene una ilusión, dije coño, para que analices la gilipollez de la edad. Dije, pues ahí está, estos se quedará en nada. ð Mira el campeón de mi Padre hay. ð Esto se quedará en nada. Se quedará en una noticia más y fin. Pasaba un poco de refilón. No es que fuese en el titular absoluto. ð Y lo comentaste con tu Padre. ð Si, lo comenté. Le dije en visto que esto no va ningún lado, estuvo son difamaciones, esto son tonterías, esto no va a nada.

Como se puede observar, el texto tiene muchos errores. La clasificación de estos es la siguiente:

Errores de reconocimiento

En este apartado, encontramos errores como cambios de número de las palabras, como es el caso de *estos*. También se incluye en esta categoría la adición de palabras nuevas, como es el caso de la conjunción *y* en la primera frase. Las sustituciones de palabras, como es el caso del uso de *que* en lugar de la preposición *de* que es la que se usa en esta expresión o el cambio del tiempo verbal, que pasa de *estábamos* a *estamos*. Y, finalmente, las omisiones como es el caso del artículo después de la preposición *de*, en el segundo párrafo, o la omisión de la preposición *a* antes del determinante *ningún*.

Errores de ortografía y puntuación.

Como hemos explicado antes, la puntuación de todo el texto se debe ir dictando, no sirve con la entonación que da el rehablador para que el sistema reconozca una pregunta o una exclamación y añada directamente los signos correspondientes. Es por eso por lo que faltan todas las interrogaciones y exclamaciones que debe haber en el texto. A parte de la puntuación, el sistema pone mayúscula siempre que se usa el sustantivo *padre*; en cambio, no identifica que cuando dice *periódico* en este caso se refiere al nombre propio de ese diario y no al sustantivo genérico, a pesar de que al redactar el texto se ha hecho la distinción pronunciando *la portada de El Periódico* y no pronunciando *la portada del periódico*. Y, finalmente, en lugar de reconocer la preposición *en*, escribe la preposición en inglés.

Errores graves

Como ya se ha explicado, en esta categoría se incluyen todos aquellos errores que alteren el significado del texto, por tanto, se irán comentando uno a uno. En primer lugar, el sistema ha sustituido el pronombre *tú* por el posesivo *su*, lo que hace que gramaticalmente sea incorrecto y que parezca que a la oración le falte un sustantivo. El siguiente error se produce debido a que el sistema ha sustituido *se lía* por *seguía*, que en este contexto no tiene ningún tipo de sentido y cambia lo que se dice en el original, como ha pasado con *Dragon*.

A continuación, confunde el sustantivo *ADE* por la preposición *ante*, que, como ocurre con el caso anterior, no tiene sentido; este error se produce porque el sistema no reconoce la palabra *ADE* como tal y lo sustituye por aquello que le resulta más familiar. En el caso del siguiente error, se clasifica como grave porque al cambiar el tiempo verbal del verbo cambia el sentido de la oración, ya que el uso de ese tiempo verbal en este caso no acontece. También, ha cambiado el verbo *eran* por el sustantivo *Irán*, que no tienen nada que ver.

Con el siguiente error pasa lo mismo que con la palabra *ADE*, el sistema no reconoce aquello que dice el locutor y lo sustituye por aquello que le resulta más parecido según su base de datos, que en

este caso ha sido cambiar *Gürtel* por *muster*. El sistema también ha confundido el determinante *lo* por el adverbio de negación *no*.

El siguiente error es de los más visibles, dado que el uso del sustantivo *ministro* en este caso no aporta nada a la oración y no tiene ningún sentido, a diferencia de lo que dice el original: *me hizo*. El sistema también ha cambiado *a ver* por *abrir*. En el siguiente caso, el sistema ha sustituido, *pero* por *tiene*, es decir, una conjunción por un verbo.

A continuación, el sistema ha sustituido el adverbio *ahí* por el verbo *hay*, que no tiene ningún sentido al final de una oración como la del ejemplo. También, sustituye el verbo haber conjugado para formar el pretérito perfecto compuesto por la preposición *en*, que no funciona delante de un participio. Y, finalmente, ha sustituido el pronombre *esto* por el verbo *estuvo*.

Como se puede comprobar tras analizar los ejemplos, la mayoría de errores que ha cometido el sistema son graves, dado que ha hecho cambios que afectan al significado del texto y cambian su sentido original.

En el caso del texto en inglés, este es el resultado obtenido tras su dictado con el sistema de *Windows*:

Texto original	Texto dictado
ð How are you? ð Good. ð Are you having a fun day? ð Yeah. ð Yeah. I'm officer Nice to meet you. Have you ever met a police officer before? ð Yeah. ð Have you ever taken a lie detector before? You have. Ok, so you're gonna be an old handed on this. So, the truth fairy is going to put these things on your fingers. He's Now we're gonna put	ð How are you ð Good ð Are you having the time of the ð Year ð I'm officer and jean. He's very nice to meet you. Have you ever minutes of police officer me from. ð The year ð Have you ever taken a learning detector test before. You have. Ok, so you are gonna be and all of Henry on this. So, no trust series easy going to put these things on your fingers. He he's are no more use up. Now we are reporting to avoid

Have you ever worn a helmet before? Like Ironman, right? Yeah you some questions and all you have to do is tell me the truth, ok? ð Yeah. ð You know what the truth is? ð Yeah. ð Ok, perfect then. What is your name? ð Eric. ð Eric, how old are you? ð Four. ð Do you go to school? ð Yeah. ð Ok. Do you like school? ð Yeah.	these Hamed on you. Have you ever worn a helmet need for. Like all men, right. I ô m going to ask you some questions. All you have to into tell me the truth. ð The year. ð Now you know what their trucks is. ð The year ð Ok, politics then. What ôs your ð Eric ð Eric. How always are you ð Four ð Do you go to school. ð Yet ð Ok, then the light school ð Year

Al igual que el texto en español, el dictado del texto en inglés también tiene muchos errores.

Errores de reconocimiento

En este caso, encontramos errores como la sustitución de *it's* por *he's*; aquí cambia la oración, pero mantiene el sentido de esta, dado que se puede interpretar fácilmente a que se refiere. A continuación, encontramos que se ha omitido la palabra *good*, pero no conlleva un error grave, dado que no cambia ni el sentido ni el significado de la oración.

El siguiente error es la repetición del sujeto *he*. Los dos siguientes errores son el mismo, la omisión de la conjunción *and*, que no conlleva un cambio grave en la oración. Y, finalmente, el sistema ha añadido la palabra *now* al inicio de la oración, pero no aporta nada. En este caso el editor del subtítulo podría borrarla.

Errores de ortografía y puntuación

A diferencia del texto que se ha dictado en español, en el texto inglés no encontramos errores de ortografía. Pero sí que ocurre lo

mismo que en el otro ejemplo, la puntuación no se corresponde a la entonación del texto original. Se deben ir dictando todos los signos de puntuación. También se podría destacar que, si se dictan contracciones, el sistema las escribirá tal cual, sin tener en cuenta que por escrito es mejor omitirlas.

Errores graves

En este texto, este tipo de errores son los que más predominan. El sistema ha reconocido mal muchos términos que han hecho que se cambie el sentido de la oración y pierda su significado original.

El primer error grave que encontramos es el cambio de *fun day* por *the time of the*, que no tiene ningún sentido. A continuación, el sistema sustituye en más de una ocasión *yeah* por otras palabras como *year* o *yet*. También confunde el nombre propio *Jimmy* con *jean*.

Al igual que con *yeah*, el sistema no reconoce bien *before* y lo sustituye por construcciones como *me from* o *need for*. También ha sustituido *met a* por *minutes of*, *lie* por *learning* y *old handed* por *all of Henry*, que no tiene nada que ver.

Los dos siguientes errores son de los más visibles, dado que se cambia la oración casi al completo. En primer lugar, ha sustituido *the truth fairy is* por *no trust series easy*. Y, a continuación, *he's gonna hook you up* por *he's not are no more use up*, que no tiene nada que ver con el original. En estos dos casos, no solo ha cambiado el sentido y el significado de la oración, si no que ha creado oraciones gramaticalmente incorrectas.

También, ha sustituido *going to put* por *reporting to avoid*, que no se asemeja en nada, salvo por la construcción *–ing* más infinitivo; *this helmet* lo ha cambiado por *these Hamed*. En el caso del cambio de *this* a *these* se podría clasificar como un error de reconocimiento, pero al acompañar a la palabra *helmet* y que esta también la sustituya, se considera error grave. El caso del siguiente error, que modifica *Ironman* por *all men*, es claramente un error grave porque el significado de la oración se modifica, pero se debe a que el sistema no reconoce

Ironman como un nombre propio y lo sustituye por aquello que fonéticamente se le asemeja más.

El sistema también ha sustituido *to do is* por *to into he's*, *the truth* por *their trucks*, *perfect* por *politics* y *old* por *always*. Finalmente, otro error que llama la atención es la sustitución de *do you like* por *then the light*, que no tiene ningún sentido en este caso.

Como hemos podido comprobar, al igual que con el español, el sistema de reconocimiento de voz de *Windows* comete muchos errores, pero los que más predominan son los errores graves. En ambos idiomas, en muchas ocasiones sustituye una palabra por otra que conlleva un cambio de sentido o de significado de la oración original.

4.3. Mejoras para los sistemas de reconocimiento del habla

Tras llevar a cabo el análisis de los errores de dos textos, uno en español y otro en inglés, y en dos sistemas distintos de reconocimiento del habla, podemos destacar cuáles son las medidas para mejorar estos sistemas. Las mejoras se diferenciaran entre aquellas que sean generales para mejorar el sistema en sí y aquellas que sirvan para mejorar el sistema cuando se quiera combinar con un editor de subtítulos para programas en directo o sin guion.

4.3.1. Mejoras de carácter general

Como hemos explicado anteriormente, nos referimos a mejoras de carácter general cuando hablamos de todo aquello que se puede hacer para mejorar aspectos como el reconocimiento de más términos, etc.

En primer lugar, hay que destacar que un buen entrenamiento del sistema es necesario para que este funcione correctamente. Como más textos dictemos y más reconozca la voz del interlocutor mejor funcionará. Es por eso por lo que, al probar el sistema con solo el entrenamiento básico u obligatorio del programa, el sistema deja de reconocer muchos términos o no acaba de reconocer del todo bien que es lo que se está dictando.

Otro de los aspectos a tener en cuenta como mejora es la opción que da *Dragon* de añadir vocabulario en cualquier momento al sistema. Esto

ayudará a todos aquellos que usen el sistema de reconocimiento del habla para dictar o para usar programas con comandos concretos de un ámbito específico, como podría ser la medicina. Además, puedes añadir ese vocabulario de distinta manera, ya sea mediante un glosario que ya se tenga preparado o manualmente.

Y, finalmente, la mejora que debería implementarse es la creación de un perfil concreto, como ya hace *Dragon.* El hecho de tener un perfil propio hará que el sistema reconozca mejor la voz del interlocutor sin que tenga interferencias de otras intervenciones. Si el sistema tiene un perfil para cada uno, se entrenará según las necesidades específicas de cada interlocutor y se adaptará al tono y la locución de cada persona de manera específica. Todo esto, contribuirá a una mejora del reconocimiento de los términos, ya que cada persona puede pronunciar de una manera u otra. Y también, permitirá la adaptación a distintos dialectos dentro de las lenguas.

Estos tres aspectos, que son de carácter general, pueden favorecer a que el sistema sea más eficaz. En el caso de *Dragon*, el hecho de que ya incorpore dos de estos aspectos hace que sea más eficaz, como se explicará a continuación en el apartado de conclusiones del trabajo.

4.3.2. Mejoras para la combinación con un editor de subtítulos

Como ya se anuncia en el título de este trabajo, y de este apartado, la idea de analizar los dos sistemas era para su combinación con un editor de subtítulos. Tras comprobar los errores más frecuentes de estos, hay algunas mejoras que se podrían introducir para que, cuando se combinara con el editor, la persona al mando de lanzar esos subtítulos no tuviera que corregir ni adaptar en gran medida el texto.

Una de las propuestas de mejora en este aspecto sería la aceptación de contracciones en inglés. Al tratarse de subtítulos, debemos tener siempre en cuenta que el espacio para el texto es reducido y debe seguir los parámetros establecidos de la empresa para la que se trabaje. Cuando a la hora de dictar, el sistema escriba *I am* en lugar de *I'm*, está añadiendo unos caracteres necesarios para otras palabras o simplemente para que ese subtítulo no ocupe tanto. Dado que los subtítulos son, mayoritariamente, la representación escrita

del parlamento que aparece en el contenido audiovisual en que se integren, si este se trata de una conversación informal, también se deberían poder representar los coloquialismos de manera escrita, aunque no sean correctos en ámbitos formales.

Finalmente, la propuesta de mejora más importante según nuestro criterio, es la opción de poder adaptar el sistema de reconocimiento del habla a los parámetros que se aplican de pautaje en cada caso. Es decir, que cuando se esté dictando, el sistema de reconocimiento del habla reconozca cuando se ha llegado al límite de caracteres por línea y de paso a la siguiente, intentando adaptar el parlamento a dos líneas por intervención. Esto ayudaría en gran parte a que se retrasara menos el tiempo de salida de los subtítulos, ya que no sería necesaria ninguna intervención del interlocutor, ya sea adaptarlo manualmente o apretando un botón cuando se crea que se ha llegado al límite.

Estos aspectos, facilitarían el trabajo de los subtituladores que deben crear e introducir los subtítulos en directo, ya que reduciría el tiempo de retraso del subtítulo. No solo sería útil para los subtítulos en directo, sino también para todos aquellos programas de los que se carece de un guion, como los programas de entrevistas, etc. El hecho de usar un sistema de reconocimiento del habla reduciría el tiempo empleado para transcribir el texto.

5. Conclusiones

Tras realizar el análisis de los dos sistemas de reconocimiento del habla podemos afirmar que la mayoría de errores que se han detectado han sido errores de reconocimiento pero graves; es decir, errores que afectaban al sentido o al significado de la oración. En el caso del texto en español encontramos errores como nombres propios o cambios de tiempos verbales que afectan a la oración. Respecto al inglés, nos encontramos con errores de mal reconocimiento de expresiones o también cambios de formas verbales.

Al comparar los dos sistema y haber realizado el mismo entrenamiento para cada uno, es decir el básico, podemos afirmar que *Dragon NaturallySpeaking* da mejores resultados que *Windows Speech Recognition*. A pesar de que Microsoft empezó antes a desarrollar su producto, desde su aparición, *Dragon* ha dado mejores resultados y es de las herramientas más usadas en este ámbito.

A pesar de solo haber usado los sistemas con su entrenamiento básico, esto nos ha permitido destacar algunos factores a tener en cuenta a la hora de usar un sistema de reconocimiento del habla. El primero, como ya se ha destacado a lo largo de todo el trabajo es realizar un buen entrenamiento de este, no solo para que reconozca nuestra voz, sino también para que mejore y sea capaz de añadir más palabras y más construcciones al sistema.

Por otro lado, también destacar la importancia de tener un buen entorno y un buen equipo con el que trabajar con el sistema de reconocimiento del habla. El uso de un buen micrófono es primordial para este tipo de trabajos. Al igual que tener un entorno sin ruidos y que permita que el dictado se vea afectado por otros sonidos. Y, finalmente, un dato que nos ha parecido relevante es tener un glosario que añadir, en el caso de *Dragon*, o con el que entrenar el sistema, para que este sea capaz de reconocer la mayoría de lo que se está dictando sin necesidad de corregirlo o modificarlo manualmente.

En cuanto a la combinación de los sistemas con un editor de subtítulos también nos ha permitido destacar algunos aspectos a tener en cuenta. El primero de estos es el hecho de tener en cuenta el pautaje del texto que estamos dictando, es decir, tener en cuenta los caracteres y el máximo de líneas en pantalla, de manera aproximada, para no tener que cambiarlo manualmente.

También deberemos asegurarnos de dictar la puntuación de las oraciones para no tener que modificarlo antes de lanzar el subtítulo al directo. Y, por último, también para no intervenir manualmente, pronunciar el texto lo más correctamente posible y así evitarnos errores que se deban corregir.

En conclusión, el uso de un sistema de reconocimiento del habla es muy útil siempre y cuando se entrene de manera correcta, para que funcione lo mejor posible, y se tenga en cuenta la finalidad del texto que se está dictando, como en este caso para subtitular, para poder obtener un buen resultado.

6. Bibliografía

Bueno, V. F., González, I., & Ruiz, B. (2007) Subtitulado en tiempo real. Sistemas y tecnología. *Discapacidad, Accesibilidad a los medios audiovisuales para personas con discapacidad*. AMADIS, 6.

Jimmy Kimmel Live (21 de junio, 2012) Jimmy Kimmel Lie Detective #2 [Archivo de video] Recuperado de https://www.youtube.com/watch?v=ZlgAirxONLo

Lago, M. D. G. (2013) Análisis de la precisión en los subtítulos en directo emitidos por RTVE. *Traducción multimedia: diversas pantallas, enfoques diversos*, 39.

Mediaset España (18 de marzo, 2019) *Willy Bárcenas en CHESTER: 'Reconocí la letra de mi padre en los papeles'* | *Mediaset España* [Archivo de vídeo] Recuperado de https://www.youtube.com/watch?v=Aiw6x5T5m6g

MUSA project (2004). MUSA project. (en línea) Sifnos.ilsp.gr. Disponible en: http://sifnos.ilsp.gr/musa/demos.html (Fecha consulta: 1 Mar. 2019).

Orero, P. (2006) Real-time subtitling in Spain. *inTRAlinea, edición especial.* Disponible en: http://www.intralinea.org/specials/article/Real-time_subtitling_in_Spain (Fecha consulta: 1 Abr. 2019).

Raud, H. F., & Brennan, P. M. (2000). U.S. Patent No. 6,125,341. Washington, DC: U.S. Patent and Trademark Office.

Rodríguez, A. P. (2005). El subtitulado para sordos: estado de la cuestión en España. *Quaderns: Revista de traducció,* (12), 161-172.

Romero-Fresco, P. (2011). *Subtitling through speech recognition: Respeaking*. St. Jerome Pub..

VOICE Project (2004). *VOICE Project's HomePage (EN)*. (en línea) Voiceproject.eu. Available at: http://www.voiceproject.eu/_voice_en.htm (Fecha consulta: 1 Mar. 2019).